DOCUMENT RESUME

ED 467 817 TM 034 361

AUTHOR	Reese, Lynda M.
TITLE	Impact of Local Item Dependence on Item Response Theory Scoring in CAT. Law School Admission Council Computerized Testing Report. LSAC Research Report Series.
INSTITUTION	Law School Admission Council, Princeton, NJ.
REPORT NO	LSAC-R-98-08
PUB DATE	1999-08-00
NOTE	17p.
PUB TYPE	Reports - Research (143)
EDRS PRICE	EDRS Price MF01/PC01 Plus Postage.
DESCRIPTORS	College Entrance Examinations; *Item Response Theory; Law Schools; Scoring; Simulation; *Test Items
IDENTIFIERS	Item Dependence; Law School Admission Test; *Local Independence (Tests)

ABSTRACT

This study represented a first attempt to evaluate the impact of local item dependence (LID) for Item Response Theory (IRT) scoring in computerized adaptive testing (CAT). The most basic CAT design and a simplified design for simulating CAT item pools with varying degrees of LID were applied. A data generation method that allows the LID among items to be defined was applied to generate five CAT item pools. The LID exhibited by these pools ranged from zero LID (complete local item independence) to extreme LID. Results indicate that for certain types of scoring an extreme amount of LID may adversely impact the final score obtained by the examinee. The estimated precision of the test was also affected by the extreme LID level studied here. For the medium level of LID, structured to display the amount of LID typically displayed by the Law School Admission Test, the effects of the LID were not troublesome. (Contains 8 figures and 36 references.) (SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 467 817

TM

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
J. VASELECK
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
1

Impact of Local Item Dependence on Item Response Theory Scoring in CAT

Lynda M. Reese
Law School Admission Council

Law School Admission Council
Computerized Testing Report 98-08
August 1999

LAW Services

A Publication of the Law School Admission Council

TM034361

BEST COPY AVAILABLE

The Law School Admission Council is a nonprofit corporation that provides services to the legal education community. Its members are 196 law schools in the United States and Canada.

Copyright© 1999 by Law School Admission Council, Inc.

All rights reserved. This book may not be reproduced or transmitted, in whole or in part, by any means, electronic or mechanical, including photocopying, recording, or by any information storage and retrieval system, without permission of the publisher. For information, write: Communications, Law School Admission Council, Box 40, 661 Penn Street, Newtown, PA 18940-0040.

LSAT® and the Law Services logo are registered marks of the Law School Admission Council, Inc.

This study is published and distributed by the Law School Admission Council (LSAC). The opinions and conclusions contained in this report are those of the author and do not necessarily reflect the position or policy of the Law School Admission Council.

Table of Contents

Executive Summary 1

Introduction 1

Methodology 2
Generation of CAT Pools 2
CAT Simulation 2
Analyses 3

Results 4
θ and Estimated True Score 4
Conditional Standard Error of Measurement 7

Discussion 9
Future Research 10

References 11

Executive Summary

In computerized adaptive testing (CAT) an attempt is made to select items for individual test takers that are appropriate for their ability level. This adaptation of the difficulty level of the test to the ability level of the test taker is made possible through the application of item response theory (IRT). IRT is a mathematical model that relates the probability that a test taker will answer a single test item (i.e., test question) correctly to the ability level of the test taker and specific characteristics of the test item. In applying IRT, a formal assumption of local item independence is made. This assumption states that once the ability level of the test taker is accounted for, the responses of test takers to individual items on the test should be statistically independent.

In a test-taking situation, many circumstances arise that cause the local item independence assumption to be violated to some degree. For instance, if a test section is especially difficult, fatigue may adversely affect the performance of test takers on the items at the end of the section. In this case, the difficulty level of the items found at the beginning of the section affect performance on later items, and so these items are said to exhibit some degree of local item dependence (LID).

The impact of LID on various applications of IRT within the paper-and-pencil mode of testing has been evaluated. Depending on the particular test design, a computerized test may rely more heavily upon IRT for such procedures as item selection and ability estimation, and so the assumptions of the model become even more important. This study represents a first evaluation of the impact of LID for IRT scoring in CAT. As such, the most basic CAT design and a simplified design for simulating CAT item pools with various degrees of LID were applied. The results indicate that, for certain types of scoring, an extreme amount of LID may adversely impact the final score attained by the examinee (i.e., test taker). The estimated precision of the test was also affected by the extreme LID level studied here. For the medium level of LID, structured to display the amount of LID typically displayed by the LSAT, the effects of the LID were not troublesome.

Future research in this area should focus on some of the computerized testing designs that are currently being evaluated for the LSAT. Also, future research should be carried out to evaluate LID levels that represent situations likely to arise in building an item pool for computerized testing. For example, the effect of 100 items displaying an extreme level of LID within a medium LID CAT pool should be evaluated.

Introduction

The local item independence assumption of item response theory (IRT) has generated a great deal of interest among psychometric researchers (Ackerman & Spray, 1987; Ackerman, 1987; Andrich, 1978, 1985; Bell, Pattison, & Withers, 1988; Chen & Thissen, 1997; Embretson, 1984; Goldstein, 1980; Jannarone, 1986, 1987, 1991a, 1991b, 1991c, 1994; Kelderman, 1984; Kempf, 1977; Lord, 1953; Masters, 1988; Pashley & Reese, 1999; Reese, 1995; Reese & Pashley, 1999; Rosenbaum, 1984; Spray & Ackerman, 1987; van den Wollenberg, 1982; Wainer & Kiely, 1987; Yen, 1984, 1993). Up to this point, however, research in this area has focused primarily on the paper-and-pencil mode of testing. While some have alluded to the fact that local item dependence (LID) is an issue that will need to be addressed within the computerized adaptive testing (CAT) environment, little research has been carried out toward this end (one exception to this is Spray, Parshall, & Thomas, 1997). One reassurance that has been cited with regard to the impact of LID within the paper-and-pencil test is that the effects are somewhat equalized across examinees, since all are asked to respond to the same set of items. The effects of LID within the CAT environment are somewhat more troublesome in that examinees (i.e., test takers) respond to different sets of items, and any adverse impact may not be equalized across examinees for this testing mode.

This research represents an initial study investigating the effects of LID within a CAT environment. A data generation method that allows the LID among items to be defined was applied to generate five CAT item pools. The LID exhibited by these item pools ranged from zero LID (complete local item independence) to extreme LID. The effect of the various levels of LID on scoring outcomes within a standard maximum information CAT were then evaluated.

The author wishes to acknowledge the valuable advice of Peter Pashley and Louis Roussos, the technical support of Jennifer Lawlor, and the support of Deborah Schnipke in providing the program used to carry out these computerized adaptive testing simulations.

Methodology

Generation of CAT Pools

Five CAT item pools consisting of 500 items each were generated. These item pools were created to display zero, low, medium, high, and extreme levels of LID. These levels of LID were defined based on analysis of actual test data as described in Reese (1995). The zero LID level serves as the baseline for comparison, the low LID level represents the lowest degree of LID observed in real data analyses, and the medium LID level represents the LID typically observed for the LSAT. The extreme LID level represents the level of LID observed for a highly dependent item set, and the high LID level was defined to fall exactly halfway between the medium and extreme levels. These LID levels have been studied with regard to paper-and-pencil testing, and the impact on various IRT outcomes in the paper-and-pencil testing environment has been evaluated (see Reese, 1995). Simulating similar LID levels for the CAT pools will allow comparisons between these two modes of testing with regard to the issue of LID.

In building an initial item pool for a CAT, testing programs are often forced to rely on item parameters obtained from paper-and-pencil pretest calibrations. Therefore, item pools were generated for this study in a way that emulates this process. Item responses were generated and calibrated for blocks of 100 items, a reasonable number of items to yield from a typical paper-and-pencil test administration. Typical LSAT item parameters were used as the generating item parameters. Responses were generated for 4000 standard normal ability values. The simulated data were calibrated using BILOG (386 BILOG 3, Mislevy & Bock, 1990). The default scoring method, number of quadrature points, and priors were utilized. To assure that the item and ability parameter estimates for the true and generated data were on a common scale, the *rescale* option in BILOG was applied, scaling the ability parameters from each calibration to have a mean of zero and a standard deviation of one. With the item parameters on a common scale, five blocks of items were then pooled to create a CAT item pool of 500 items for each level of LID.

The CAT item pools were generated using a method described by Pashley and Reese (in press). This data generation method is an improvement over other methods of simulating LID data in that the structure of the LID among the items is defined, rather than the relationship among multiple abilities. The steps carried out in applying this method are as follows.

1. Define a desired correlation structure among the test items.
2. Generate a vector x of multivariate random deviates according to the correlation structure defined in step 1.
3. Using the inverse normal transformation, transform x to y, a vector of dependent uniform (0,1) deviates.
4. Compare each uniform deviate to individual item correct probabilities in order to obtain 0 or 1 item responses.

CAT Simulation

For each of the five CAT item pools, responses to a fixed-length 25-item standard maximum information CAT were generated for 1,000 standard normal ability (θ) values. Items were selected for administration from the five LID item pools based on the criterion of maximum information, with item information defined by the equation

$$I_i(\theta) = \frac{2.89a_i^2(1-c_i)}{[c_i + e^{1.7a_i(\theta-b_i)}][1+e^{-1.7a_i(\theta-b_i)}]^2}, \qquad (1)$$

where i represents the item number,
a represents the IRT discrimination parameter,
b represents the IRT difficulty parameter, and
c represents the IRT lower asymptote parameter
(see, e.g., Hambleton, Swaminathan, & Rogers, 1991).

To assure that the most informative items in the pool would not become overexposed, an exposure control method described by Kingsbury and Zara (1989) was incorporated into the simulations. Applying this method, the ten items with the highest information values at $\theta = 0$ (the starting value for all simulated examinees) were identified, and the first item to be administered to a simulated examinee was randomly selected from among these ten items. The second item was randomly selected from the nine best items at the new estimate of θ. The third item was randomly selected from the eight best items, and so on until, beginning with the tenth item, the item with the highest information was selected. At each step of this process, if the selected item had already been administered to the simulated examinee, the next best item was selected.

After each item was selected, the simulated examinee's score (right/wrong) on the item was determined using the true item parameters and their true θ. Their θ estimate was then updated based on the LID item parameters using Owen's Bayesian sequential scoring (Owen, 1969). After all items were administered, a Bayesian modal score (e.g., see Hambleton, Swaminathan, & Rogers, 1991) based on a standard normal prior was calculated and used as the final θ estimate.

In addition to ability estimates, an estimated true score was determined for each simulated examinee at each LID level. Since each simulated examinee responded to a different set of items in the CAT simulation, an estimated true score on a common set of items was determined for each simulated examinee. This technique has been applied for some operational adaptive tests (Stocking, 1996). Here, for each simulated examinee's θ estimate derived from the CAT simulation, the corresponding estimated true score on a 101-item base LSAT form was determined for each LID level. The estimated true score corresponding to each simulated examinee's true θ value was also determined and treated as the true value in evaluating the estimated true scores. In this way, the estimated true scores for different simulated examinees at different LID levels could be compared.

For each ability estimate derived for each LID level, an estimate of the conditional standard error of measurement ($CSEM_E$,where the subscript E indicates that this is an estimate) was calculated. This was accomplished by using the item parameters from the LID item pools and applying Equation 1 to calculate the item information at the final ability estimate for each item administered to a simulated examinee. Item information was then summed across the items administered to a simulated examinee and the square root of the reciprocal of this value was used as the $CSEM_E$. In order to evaluate the $CSEM_E$ values, the *CSEM* was also derived for each ability estimate by applying this same method described above using the true item parameters for those items administered to the simulated examinee at each LID level. While this *CSEM* is not a "true" *CSEM* in the strictest sense, it is denoted as $CSEM_T$ to indicate that the true item parameters were used in the calculation.

Analyses

Root mean squared error. The root mean squared error (*RMSE*) between the true and estimated ability values was studied at each level of LID in order to determine the degree to which the ability estimates depart from the true ability parameters. The θ estimates were first grouped based on true values of the θ parameter. Within each of these groupings, *RMSE* was calculated by applying the equation

$$RMSE = \left[\frac{1}{n}\sum_{i=1}^{n}(true_i - est_i)^2\right]^{1/2}, \qquad (2)$$

where n is the number of simulated examinees,
$true_i$ represents the true θ-parameter for examinee i, and
est_i represents the θ-parameter estimate for examinee i.

Equation 2 was also applied to calculate the *RMSE* statistic comparing the estimated true scores derived for the five LID levels to the estimated true scores determined for the true θ parameters.

Root mean squared difference. A root mean squared difference (*RMSD*) statistic was studied to compare the $CSEM_E$ values to the $CSEM_T$ values for each LID level. Simulated examinees were also grouped by their true θ values for these analyses. The equation for this statistic is given by

$$RMSD = [\frac{1}{n}\sum_{i=1}^{n}(CSEM_T - CSEM_E)^2]^{1/2}. \quad (3)$$

Bias statistic: The *bias* statistic comparing the ability estimates and estimated true scores was calculated in a similar manner to the *RMSE* by applying the equation

$$Bias = \frac{1}{n}\sum_{i=1}^{n}(true - est). \quad (4)$$

The bias in SEM_E relative to $CSEM_T$ was studied by applying the equation

$$Bias = \frac{1}{n}\sum_{i=1}^{n}(CSEM_T - CSEM_E). \quad (5)$$

Like the *RMSE* and *RMSD* analyses, simulated examinees were grouped by their true θ values in these bias analyses.

Results

θ and Estimated True Score

The *RMSE* and *bias* analyses of the ability estimate for each LID level are presented in Figures 1 and 2, respectively. Note that for these figures and all subsequent figures of their kind, the x-axis points represent the midpoint of a range of θ values. Figure 1 indicates that θ was estimated with roughly equal precision at the center of the ability scale for the five LID levels studied. However, for ability values lower than –1 and higher than .5, the high and extreme LID levels begin to depart from the other LID levels, displaying more error. Figure 2 reveals that while there is a slight tendency at the zero through medium LID levels to overestimate low scores and underestimate high scores (a consequence of Bayes modal scoring), this effect is intensified for the high and extreme LID levels.

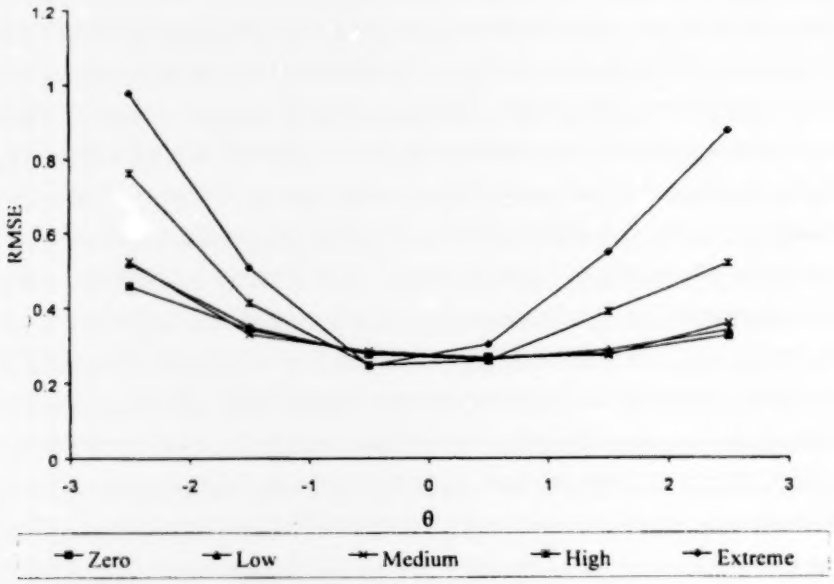

FIGURE 1. *RMSE statistic for θ estimates grouped by true θ parameters*

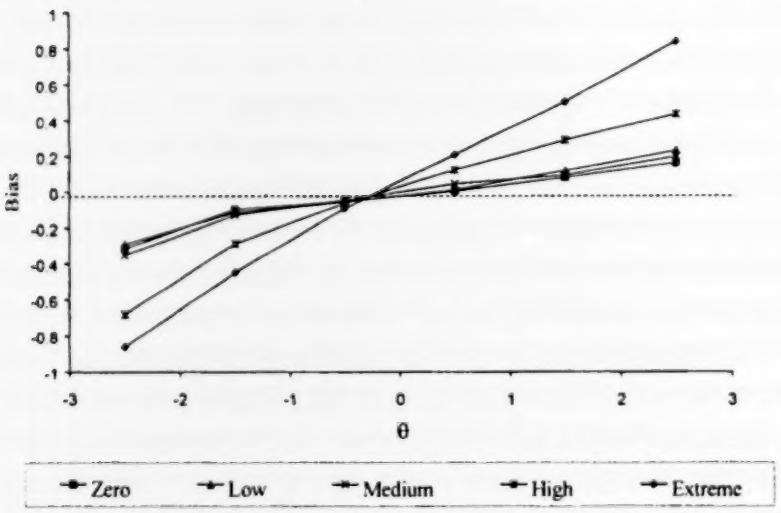

FIGURE 2. *Bias statistic for θ estimates grouped by true θ parameters*

BEST COPY AVAILABLE

Figures 3 and 4 present the results for the estimated true score for each LID level. Recall that the estimated true scores were derived by converting the θ estimates obtained from the CAT simulations to an estimated true score on a 101-item base LSAT paper-and-pencil form. These figures reveal the same pattern of results observed for the θ scale. This is to be expected given the way in which these estimated true scores were derived.

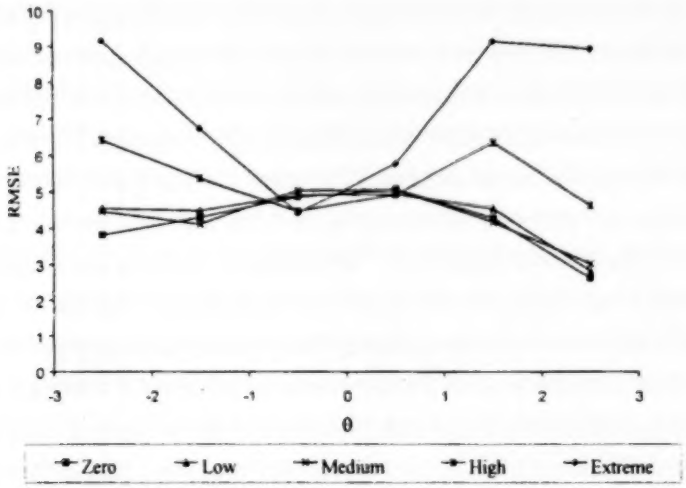

FIGURE 3. *RMSE statistic for the estimated true score grouped by true θ parameters*

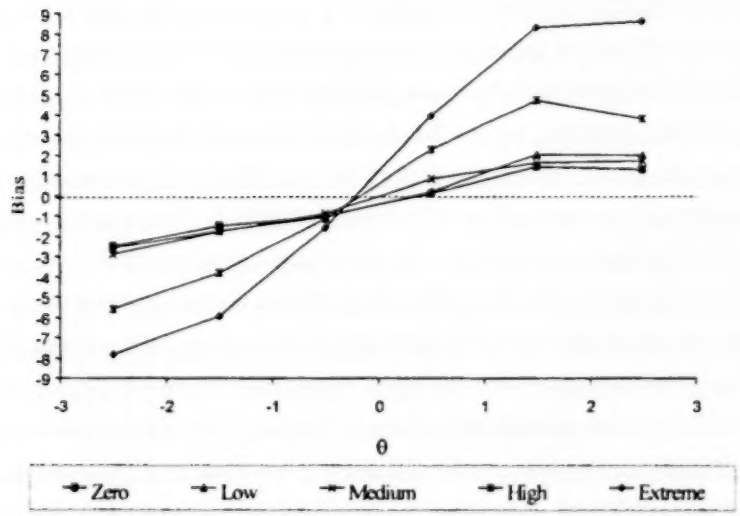

FIGURE 4. *Bias statistic for the estimated true score grouped by true θ parameters*

Conditional Standard Error of Measurement

While the precision of the θ estimates and estimated true scores are important outcomes in CAT, our estimate of the precision with which we are estimating ability throughout the testing session is also important. For this

reason, the impact of the LID induced here on the estimation of the *CSEM* was evaluated. The *RMSD* analyses displayed in Figure 5 indicate that the zero, low, and medium LID levels display a similar amount of error in *CSEM* estimation for the entire θ scale. For the high and extreme LID levels, the *CSEM* is estimated less precisely for most of the range of the θ scale. One exception to this is the high LID level at the lowest extreme of the ability scale. It would appear that the *CSEM* is estimated with more precision at the high LID level than at the other four LID levels for this portion of the ability scale. There are few examinees in this region of the ability scale, so this result is probably due to some instability in the estimation of this measure. *Bias* analyses presented in Figure 6 show a tendency toward underestimating the *CSEM* for the extreme LID level. With the exception of the highest extreme of the ability scale, the high LID level also shows a tendency toward underestimating the *CSEM*, but to a lesser degree than the extreme LID level. These results suggest that our assessment of the precision of the θ estimate is impacted by the LID induced here.

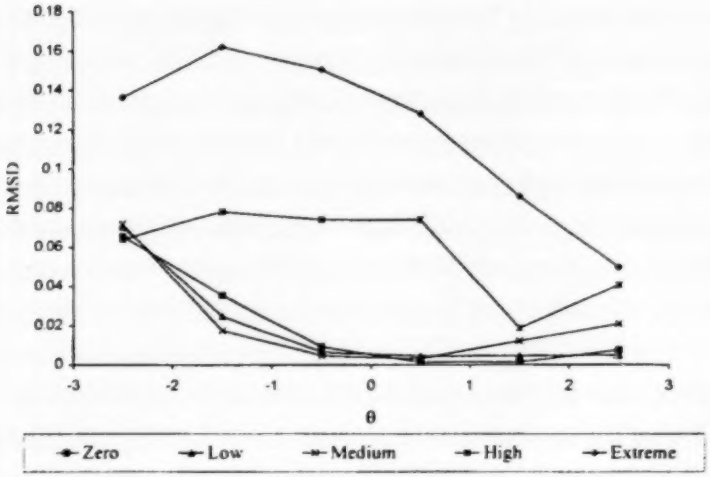

FIGURE 5. *RMSD statistic for the CSEM grouped by true θ parameters*

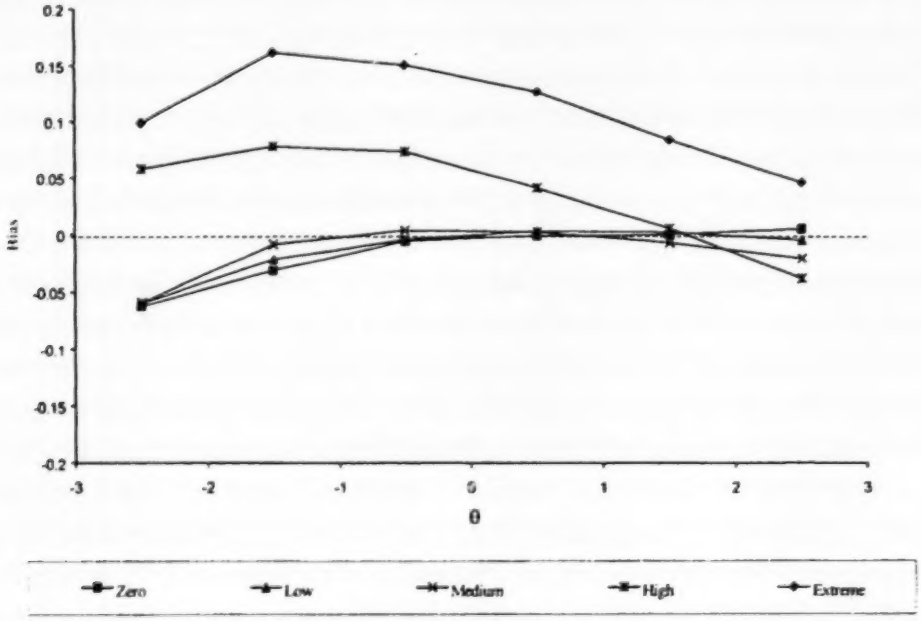

FIGURE 6. *Bias statistic for the CSEM grouped by the true θ parameters*

Discussion

The results of this study indicate that ability estimation is adversely impacted by the high and extreme LID levels simulated here. Lower values of θ are overestimated and higher values of θ are underestimated. Figure 7, taken from Reese (1995), provides some explanation for this result. This figure overlays test characteristic curves for a 101-item paper-and-pencil test for the zero and extreme LID levels simulated for this study. Since the LID causes the c-parameter values to be underestimated and the a-parameter values to be inflated, a steeper test characteristic curve is produced. The steepness of this curve results in more θ estimates being concentrated in the center of the ability scale. This effect resulted in the bias of θ estimation presented in Figure 2.

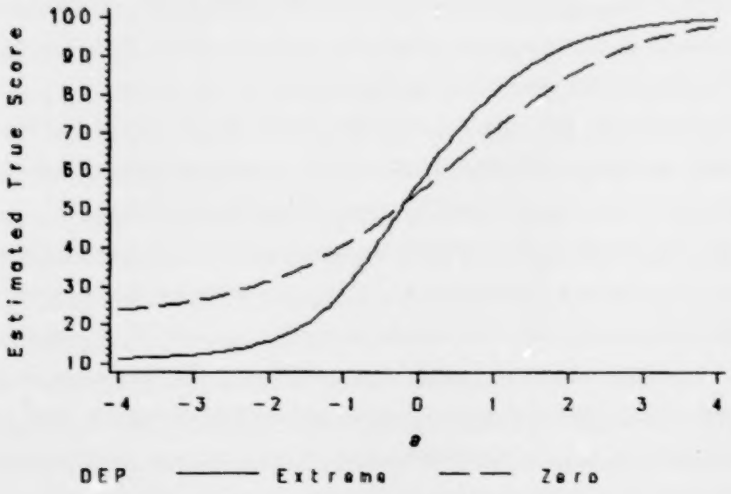

FIGURE 7. *Test characteristic curve overlay plot for the zero and extreme LID levels*

While θ estimates were studied here, it is unlikely that θ estimates would be reported to examinees since this measure has very little meaning to them. For this reason, θ estimates were converted to an estimated true score on a base form, and this scoring outcome was evaluated. The results observed for the estimated true score were very similar to those results observed for θ. This is to be expected given the way in which these estimates were derived. Figure 7 reveals that the effect on the estimated true score as it is calculated from the item parameters is quite different from the results observed here. However, directly calculating the estimated true score would not be appropriate within a testing environment wherein examinees respond to different sets of items.

In addition to ability estimation, the impact of LID on the estimation of the *CSEM* was studied here. The results presented indicate that the *CSEM* tends to be underestimated for much of the ability scale for the high and extreme LID levels. Based on previous research (Reese, 1995; Yen, 1993) this result is not surprising. As Figure 8, taken from Reese (1995), shows, the extreme LID level causes information to be overestimated for much of the ability scale. Since the *CSEM* is the reciprocal of information, the under-estimation of this measure is expected. At the highest extreme of the ability scale, Figure 8 shows that there is a slight tendency for information to be underestimated, which may account for the results observed here for the high LID level. A fixed-length adaptive test was simulated for this study, but the impact of this *CSEM* under-estimation would be problematic within a variable-length test wherein the *CSEM* could be used as the termination criterion. For a high and extreme level of LID, testing could be terminated too soon since the precision of the test would be overestimated. In itself, the over-estimation of information could result in the examinee responding to a different set of items than would have been selected with locally independent data.

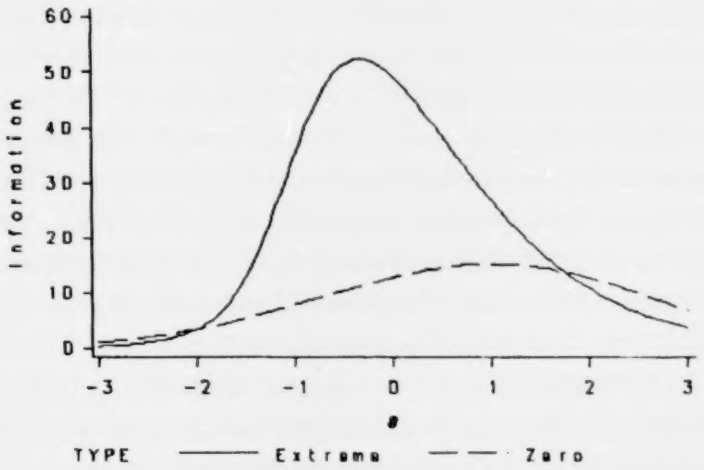

FIGURE 8. *Test information function overlay plot for the zero and extreme LID levels*

As stated earlier, the LID levels studied here were defined based on analyses of actual test data. The medium LID level was defined to represent the level of LID typically displayed by the paper-and-pencil LSAT, and the effects of the LID for this level were not problematic. It is only for the high and extreme LID levels, which are not as likely to occur in practice, that the results observed for this study become problematic.

Future Research

This study represents an initial investigation into the consequences of LID for CAT. As such, a standard maximum information CAT that adapted after the response to each item was simulated. However, this pure form of CAT is not practical for many testing programs. For the LSAT in particular, a CAT design that utilizes testlets seems likely (see Reese & Schnipke, 1999; Reese, Schnipke, & Luebke, 1999; and Schnipke & Reese, 1999). The impact of LID would probably vary for different test designs and scoring schemes, and this should be addressed by future research.

In addition to adhering to the most basic CAT design, the simulation of the CAT pools for this research was defined to be as simplified as possible. For each of the LID levels, each of the five blocks of 100 items that were pooled together to form the 500-item CAT pool was defined to represent an identical level of LID. While it was important to begin with this simplified data simulation design, it seems unlikely that such a situation would arise in practice. Using this study as a baseline, other designs for simulating LID CAT pools should be studied. For instance, the effect of one 100-item block of extreme LID items within a medium LID CAT pool should be evaluated.

The CAT simulation was also simplified in that item response scores (correct or incorrect) were assigned by using the true item and ability parameters. These responses were not affected by the LID of the data. In an actual testing situation in which there were local dependencies among items, the responses of the examinees to these items would be correlated to some degree. The CAT simulation carried out here did not reflect this aspect of the LID. Instead, this study evaluated the effect of an item pool contaminated by LID on the IRT scoring outcomes in CAT. Future research should address the effect of the LID on the item responses of the simulated examinees, perhaps by obtaining item responses from the initial response data created in forming the CAT item pools. Such responses would display the correlation structure that would be expected as a result of the LID.

References

Ackerman, T. (1987). The robustness of LOGIST and BILOG IRT estimation programs to violations of local independence (RR-87-14). Iowa City, IA: ACT.

Ackerman, T., & Spray, J. (1987). *A general model for item dependency* (RR-87-9). Iowa City, IA: ACT.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43*, 561-573.

Andrich, D. (1985). A latent trait model for items with response dependencies. In S. E. Embretson (Ed.), *Test design: Developments in psychology and psychometrics*. Orlando, FL: Academic Press.

Bell, R. C., Pattison, P. E., & Withers, G. P. (1988). Conditional independence in a clustered item test. *Applied Psychological Measurement, 12*, 15-26.

Chen, W., & Thissen, D. (1997). Local dependence indices for item pairs using item response theory. *Journal of Educational and Behavioral Statistics, 22*, 265-289.

Embretson (Whitely), S. (1984). A general latent trait model for response processes. *Psychometrika*, 49, 175-186.

Goldstein, H. (1980). Dimensionality, bias, independence and measurement scale problems in latent trait test score models. *British Journal of Mathematical and Statistical Psychology, 33*, 234-246.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Thousand Oaks, CA: SAGE Publications, Inc.

Jannarone, R. J. (1986). Conjunctive item response theory kernels. *Psychometrika*, 51, 357-373.

Jannarone, R. J. (1987). Locally dependent models for reflecting learning abilities (Center For Machine Intelligence Report *#87-67)*. Columbia, SC: University of South Carolina.

Jannarone, R. J. (1991a). Conjunctive measurement theory: Cognitive research prospects. In M. Wilson (Ed.), *Objective Measurement: Theory into practice, volume 1*. Norwood, NJ: Ablex Publishing Corporation.

Jannarone, R. J. (1991b). Locally dependent cognitive process measurement, contrasts and connections with traditional test theory. Unpublished Manuscript, University of South Carolina.

Jannarone, R. J. (1991c). *Measuring quickness and correctness concurrently: A conjunctive IRT approach.* Unpublished Manuscript, University of South Carolina.

Jannarone, R. J. (1994). Local dependence: Objectively measurable or objectionably abominable? In M. Wilson (Ed.), *Objective Measurement: Theory into Practice, Volume 2*. Norwood, NJ: Ablex Publishing Corporation.

Kelderman, H. (1984). Loglinear Rasch model tests. *Psychometrika, 49*, 223-245.

Kempf, W. F. (1977). A dynamic test model and its use in the microevaluation of instructional material. In H. Spada & W. F. Kempf (Eds.), *Structural models of thinking & learning*. Vienna: Hans.

Kingsbury, G. G., & Zara, A. R. (1989). Procedures for selecting items for computerized adaptive testing. *Applied Measurement in Education, 2*, 359-375.

Lord, F. M. (1953). The relation of test score to the trait underlying the test *Educational and Psychological Measurement, 13*, 517-549.

Masters, G. N. (1988). Item discrimination: When more is worse. *Journal of Educational Measurement, 25*, 15-29.

Mislevy, R. J., & Bock, R. D. (1990). *BILOG*. Mooresville, IN: Scientific Software, Inc.

Owen, R. J. (1969). *A Bayesian approach to tailored testing* (Research Report RB-69-92). Princeton, NJ: Educational Testing Service.

Pashley, P. J., & Reese, L. M. (in press). *On generating locally dependent item responses* (Statistical Report 95-04). Newtown, PA: Law School Admission Council.

Reese, L. M. (1995). *The impact of local dependencies on some LSAT outcomes* (Statistical Report 95-02). Newtown, PA: Law School Admission Council.

Reese, L. M., & Pashley, P. J. (1999). *Impact of local item dependence on true-score equating* (Research Report 97-01). Newtown, PA: Law School Admission Council.

Reese, L. M., & Schnipke, D. L. (1999). *An evaluation of a two-stage testlet design for computerized adaptive testing* (Computerized Testing Report 96-04). Newtown, PA: Law School Admission Council.

Reese, L. M., Schnipke, D. L., & Luebke, S. L. (1999). *Incorporating content constraints into a multi-stage adaptive testlet design* (Computerized Testing Report 97-02). Newtown, PA: Law School Admission Council.

Schnipke, D. L., & Reese, L. M. (1999). *A comparison of testlet-based test designs for computerized adaptive testing* (Computerized Testing Report 97-01). Newtown, PA: Law School Admission Council.

Stocking, M. L. (1996). An alternative method for scoring adaptive tests. *Journal of Educational and Behavioral Statistics, 21*, 365-389.

Rosenbaum, P. R. (1984). Testing the conditional independence and monotonicity assumptions of item response theory. *Psychometrika, 49*, 425-435.

Spray, J. A., & Ackerman, T. (1987). *The effect of item response dependency on trait or ability dimensionality* (Research Report 87-10). Iowa City, IA: ACT.

Spray, J., Parshall, C. G., & Thomas L. (1997). *Calibration of CAT items administered online for classification: assumption of local independence*. Paper presented at the annual meeting of the Psychometric Society, Gatlinburg, TN.

van den Wollenberg, A. L. (1982). Two new test statistics for the Rasch model. *Psychometrika*, 47, 123-140.

Wainer, H., & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement, 24*, 185-201.

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement, 8*(2), 125-145.

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement, 30*, 187-214.

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

NOTICE

Reproduction Basis

[X] This document is covered by a signed "Reproduction Release (Blanket)" form (on file within the ERIC system), encompassing all or classes of documents from its source organization and, therefore, does not require a "Specific Document" Release form.

[ ] This document is Federally-funded, or carries its own permission to reproduce, or is otherwise in the public domain and, therefore, may be reproduced by ERIC without a signed Reproduction Release form (either "Specific Document" or "Blanket").